Running head: DESCRIPTIVE BOOTSTRAP

Using the Descriptive Bootstrap to Evaluate Result Replicability (Because Statistical Significance Doesn't)

Sarah Spinella

Texas A&M University

Paper presented at the annual meeting of the Southwest Educational Research Association, San Antonio, February 2, 2011.

Abstract

As result replicability is essential to science and difficult to achieve through external replicability, the present paper notes the insufficiency of null hypothesis statistical significance testing (NHSST) and explains the bootstrap as a plausible alternative, with a heuristic example to illustrate the bootstrap method. The bootstrap relies on resampling with replacement from the original sample to generate a sampling distribution of bootstrap samples. An estimated standard error can be derived from the bootstrap sampling distribution to determine the replicability of statistical results in the original sample.

Using the Descriptive Bootstrap to Evaluate Result Replicability (Because Statistical Significance Doesn't)

Result replicability "is almost universally accepted as the most important criterion of genuine scientific knowledge" (Rosenthal & Rosnow, 1984, p. 9). It is paramount that scientists have replicable findings; otherwise their research cannot be relied upon and built on by others. Thompson (2006) illustrates the importance of replicability with the story of cold fusion from 1989, when two scientists reported the possibility of using cold fusion to generate cheap, almost unlimited, clean energy. After they held a press conference announcing their finding, essentially no one was able to replicate their results. Imagine the embarrassment that necessarily ensues from publishing results that are not replicable! Thompson (1993, p. 368) stated that an emphasis on replicability "is compatible with the basic purpose of science: isolating conclusions that replicate under stated conditions."

The most straightforward type of replicability is external replicability. This is achieved by conducting the same study with a different sample and finding whether results confirm the original results. However, generating additional samples for external replicability means more time and more money must be spent. As a result, many scientists mistakenly turn to null hypothesis statistical significance testing (NHSST) as evidence of replicability, for reasons that will shortly be explained. Fortuitously, another kind of replicability exists. Internal replicability can be verified through several statistical methods, including cross-validation, the jackknife, and the bootstrap. The bootstrap generates additional samples from an original sample and uses the generated samples to estimate result replicability. The present paper first explores some of the mistakes in thinking that NHSST can be used to evaluate result replicability, while focusing primarily on explaining the concepts and methodology of using the bootstrap method.

The Insufficiency of Statistical Significance Testing

Statistical significance testing is a method of making inferences about a sample from a population. For a scientist interested in researching a question, the population is everyone for whom the question is applicable. The answer to a question about the population is called a parameter. However, collecting information from every single person in a population is usually infeasible, besides being tedious, costly, and time-engulfing. So, instead, most scientists collect information only from some of the people in the population, which is called a sample. The answer to a question about a sample is called a statistic. The essence of making inferences in social science research is finding ways to use the sample statistic to answer questions about population parameters (Thompson, 2006).

For a sample to answer questions about a population, certain things must be true about the sample. The most important is that the sample must be randomly selected from the population (Efron & Tibshirani, 1993; Manly, 1994; Sprent, 1998). If the sample was drawn from only one part of the population, it is likely that the sample statistic will be biased. For example, if researchers were interested in the correlation between hair color and high school GPA, but only collected data from smart, upper-class, brown haired people, they would most likely infer that having brown hair leads to a higher high school GPA, which is false. Additionally, the sample must be of a large enough size to make generalizations about the population.

Because NHSST relies on finding the probability of the sample result in comparison to a broader population, assumptions have to be made about the broad population as well. In the above brown-hair example, researchers looking at the variable of hair color were comparing their brown-haired sample to what they assume is true about the broad population of people with all hair colors. Compared to GPAs of everyone else, they inferred that brown hair led to higher GPAs. For NHSST, researchers usually assume that the broad population has a normal distribution. This assumption applied to GPA says most people have an average GPA, and the number of people with a certain GPA decreases as the GPA gets more extreme. Employing this assumption may be problematic because it is never completely true that a population has a normal distribution, and frequently populations do not fit very well at all under the bell-curve (Lunneborg, 2000).

In NHSST the probability of the sample statistic in comparison to a broader null normal population, given the sample size, is computed as $p_{\text{calculated}}$, usually by a statistical software package. More specifically, $p_{\text{calculated}}$ is the probability that the sample statistic could have occurred randomly in a sample from a population exactly described by the null hypothesis (in other words having brown hair made no difference in GPA) with a certain size sample. $P_{\text{calculated}}$ is evaluated against arbitrarily chosen probability cut-off points known as $\mathrm{p}_{\text{critical}}$ in order to determine statistical significance. Because $p_{\text{calculated}}$ is affected by sample size, significant results can always be obtained by increasing the sample size. Therefore $p_{\text{calculated}}$ is confounded as a method of evaluating results. Also, as Thompson (2006) noted, the inference in NHSST is from the population to the sample, not vice versa. Some mistakenly infer that a small $p_{\text{calculated}}$ value means that the population parameter must be similar to their sample statistic. They build further on this mistaken inference by thinking that if their statistic is similar to the population parameter, then other samples would likely have the same results, which is thinking that NHSST can evaluate replicability (Thompson, 1996). NHSST cannot evaluate result replicability because NHSST cannot make inferences about the population from the sample (Thompson, 2006).

Bootstrapping

Bootstrapping is a viable solution to this problem of evaluating results for replicability. While the bootstrap still requires that original samples come from random sampling, it does not make any assumptions about normality that must be confirmed in the population. The bootstrap creates new samples by resampling from the original sample to test result replicability. Thus it is named in reference to pulling yourself out of the mud with your own bootstraps (Manly, 1994).

The bootstrap resamples, with replacement, from the original data to create new samples. The new samples have the same number of scores (*n*) as the original sample, but differ in that some of the original scores may not be included in a given bootstrap sample, and some may appear multiple times. This is the result of sampling with replacement. To illustrate sampling with replacement, imagine a raffle with multiple prizes. There are an equal number of prizes to the number of participants. In sampling with replacement, after Jim’s name is chosen to win a new road bike, the hat-holder folds the name again and puts it back in the hat. Jim’s name could be chosen again to win the i-pad, and again to win the sound system. In fact, Jim’s name could be chosen every single time and he could go home with all the prizes, and his peers would not win anything. Just like the raffle, the bootstrap has the same number of scores (prizes) as the original data set (participants), but almost certainly not all of the original scores will appear in a given bootstrap sample, and some will appear more than once because of sampling with replacement. Random sampling with replacement is necessary to ensure that new samples are independent of the original sample, so they can be used to evaluate replicability (Efron & Tibshirani, 1993).

In bootstrapping, multiple bootstrap samples are created from the original data. This is like creating many different scenarios of how many prizes each person won in the raffle analogy.

Jim could have won them all, or Jenny, Tom, and Martha could have each won varying numbers of prizes and everyone else none, etc. The number of possible samples (scenarios) is overwhelming for anything other than an original data set with a small *n*. The number of possibilities can be calculated by $\frac{(2n-1)!}{n!(n-1)!}$, but in most cases only a fraction of these will be used for the bootstrap samples. For example, if *n*=3 there will only be ten possible samples, and if n=4 there are 35 possible samples, but even as low as *n*=15 there are over two billion possibilities! For the descriptive bootstrap, typically 1000-5000 bootstrap samples are created (Efron & Tibshirani, 1993). This would be a lot of work by hand, which is why this method "would have been unthinkable 30 years ago" according to Diaconis and Efron in 1983 (p. 116). But computers make it easy to do bootstrapping.

Each bootstrap sample has a sample statistic computed from its scores. Whatever statistic is the statistic of interest, be it the sample mean, the sample median, or the sample coefficient of kurtosis, that statistic will be computed for each of the bootstrap samples. Then the statistics from each of these 1000-5000 bootstrap samples will be compiled as a new distribution—a sampling distribution—as the scores in this distribution come from multiple samples, not from one sample. This bootstrap sampling distribution is used to find a grand mean and the standard error. The grand mean is the mean of all the statistics found from each bootstrap sample, and the standard deviation of the 1000-5000 statistics in the bootstrap sampling distribution is the standard error (Efron & Tibshirani, 1993).

The standard error is also known as the sampling error because it predicts how much error is contained in the sample. Because the standard error is the standard deviation of the sampling distribution, and given a sufficiently large sample *n* the sampling distribution of bootstrap samples will approach a normal distribution, it can be expected that approximately 68%

of replicated studies will fall within one standard error of the grand mean (Thompson, 2006). By adding and subtracting the standard error from the grand mean, the 68% interval can be estimated.

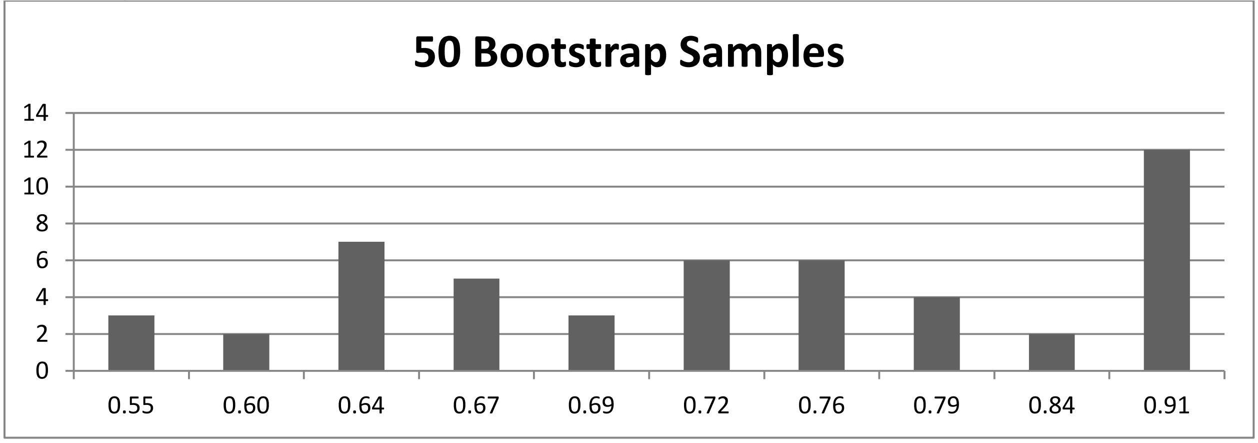
Both the grand mean and standard error resulting from a bootstrap sampling distribution are estimates. Remember that there are many possible new samples and we only used 1000-5000 of them. If different possible samples had been used, then the grand mean and the standard error would be different. According to the Central Limit Theorem, if infinitely many bootstrapping samples went into the bootstrap sampling distribution, then the grand mean and standard error would get closer and closer to certain numbers, but that kind of precision is not necessary for determining result replicability (Efron & Tibshirani, 1993). "If the mean estimate is like our sample estimate and the standard deviation of estimates from the resampling is small, then we have some indication that the result is stable over many different configurations of subjects" (Thompson, 1993, p. 373). In other words, a grand mean estimate similar to the statistic from the original sample, and a small standard error estimate mean the results are likely replicable, because the statistic stayed relatively the same even though the bootstrap samples included many different variations and combinations of the original data.

A Heuristic Example. Let's say there were three children in elementary school that you used for your sample. You were interested to see if their grades on 12 tests in the past semester correlated with the amount of sleep they got the week preceding each test. Child A had a correlation of r= .55, child B had a correlation of r= .69, and child C had a correlation of r= .91. The mean correlation for the three of them is r= .72. Even though you only have data for these three children (which is really not big enough for any statistical purposes, only a heuristic example), you are curious to know whether your classmate Katy is likely to get a similar answer to you, using three children from another school, because she wants to check answers with you before you turn in the assignment. In other words, you have an original sample of three correlation scores for children and a sample statistic, in this case a mean, based on those original scores; you want to know how replicable your results are in school children at large.

Table 1. Possible Bootstrap Samples for Example

	All possible bootstrap samples			Bootstrap sample means
1	0.91	0.91	0.91	0.91
2	0.91	0.91	0.55	0.79
3	0.55	0.55	0.91	0.67
4	0.55	0.55	0.55	0.55
5	0.55	0.55	0.69	0.60
6	0.69	0.69	0.55	0.64
7	0.69	0.69	0.69	0.69
8	0.69	0.69	0.91	0.76
9	0.91	0.91	0.69	0.84
10	0.55	0.69	0.91	0.72

Figure 1. 50 Bootstrap Samples for Example



It’s too late for you to collect more data now, because the other children didn’t keep track of their sleep, so you decide to use an internal test of replicability: the bootstrap. Given that there are only three people in your sample, there are only ten possible combinations for a bootstrap sample. They are listed in Table 1 along with their corresponding bootstrap sample means. You use a computer to take 50 random samples with replacement from your original sample. With only ten possible samples, most likely your bootstrap samples will be the same as some other bootstrap samples, though this would be unlikely to happen in a larger sample as discussed previously. When the computer takes random samples, it is choosing one of the samples in the table and then choosing another sample in the table, possibly the same one. It will generate a list of perhaps 50 samples, all from the original ten possibilities. A graph of such a list might look like Figure 1. This process might be understood more clearly by practicing running the spreadsheets found in Appendix A in Microsoft Excel with any three numbers. For the 50 samples in Figure 1, the grand mean came out as *r*= .75, but the standard error came out to .76. The grand mean is close to the sample statistic you found originally, but the standard error is so large in comparison (adding and subtracting it from the grand mean includes the whole spectrum of possible correlations) that there is very little chance your results would replicate.

Discussion

As mentioned earlier, bootstrap results are estimates, so the useful answer when using bootstrapping is not a particular grand mean, but the interval around the grand mean. A large interval compared to the range of the original scores means the results are unlikely to replicate. This could be due to a sample that is too small and makes spread out scores lend the instability of outliers to the results, or to a sample that is too variable, with too many outliers. Good (2001, p. 15) reminded us that “the bootstrap will not reproduce all the relevant characteristics of a population, only those present in the sample.” That is to say, if the original sample is missing an important section of data, or biased, then the bootstrap replications will also be similarly biased. Thompson (1993) commented that because the bootstrap samples are all from the same sample, it gives an inflated estimate of replicability, but an inflated estimate is better than not having any estimate at all. Scientists must be wary of their samples when using bootstrapping; if the original sample is not representative of the population, then they may come away with a faulty estimate of replicability.

The usefulness of the bootstrap is that it can be used to estimate the accuracy of almost any statistic. Efron and Tibshirani (1993) noted that prior to Efron developing the bootstrap, there was no easy way to estimate standard error for many statistics other than the mean. With the help of the bootstrap and computers, it is easy to estimate the replicability of results.

References

Diaconis, P., & Efron, B. (1983). Computer-intensive methods in statistics. Scientific American, 248(5), 116-130.

Efron, B., & Tibshirani, R.J. (1993). An introduction to the bootstrap. New York: Chapman and Hall.

Good, P.I. (2001). Resampling methods (2nd ed.) Boston: Birkhauser.

Lunneborg, C.E. (2000). Data analysis by resampling: Concepts and applications. Pacific Grove, CA: Duxbury.

Manly, B.F.J. (1994). Randomization and Monte Carlo methods in biology (2nd ed.). London: Chapman and Hall.

Rosenthal, R., & Rosnow, R.L. (1984). Essentials of behavioral research: Methods and data analysis. New York: McGraw-Hill.

Sprent, P. (1998). Data driven statistical methods. London: Chapman and Hall.

Thompson, B. (1993). The use of statistical significance tests in research: Bootstrap and other alternatives. Journal of Experimental Education, 61, 361-377.

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. Educational Researcher, 25(2), 26-30.

Thompson, B. (2006). Foundations of behavioral statistics: An insight-based approach. New York: Guilford.